PERCEPTUAL FEATURES FOR THE IDENTIFICATION OF ROMANCE LANGUAGES

I. Vasilescu, F. Pellegrino, J.-M. Hombert

Dynamique du Langage - I.S.H.
14, avenue Berthelot, 69363 Lyon Cedex 07 – France

ABSTRACT
This paper deals with perceptual identification and differentiation of five Romance languages, namely French, Italian, Spanish, Portuguese and Romanian. Previous studies have investigated human capability to identify spoken samples in unknown languages after a relatively brief exposure. Moreover, they have shown that subjects use perceptual categories and adapt foreign categories to their own categories in language differentiation. Accordingly, we conduct an analysis to determine which perceptual categories are salient in Romance languages identification and discrimination. Four different sets of listeners are tested. Each set consists in speakers of a different mother tongue (French, Romanian, Japanese and American English native speakers). Results reveal that identification scores are a function of the previous exposure of the listeners to the languages. Moreover, the strategies of discrimination among languages are mother tongue dependent and several potential features emerge that may be relevant in automatic language identification.

1. INTRODUCTION
Research in human identification of languages benefits from a special attention in the last years, as an alternative to improve the robustness of automatic systems for language recognition. This approach highlights the ability of human beings to identify languages and to take advantage of a short training period [3]. Even without any exposure to a foreign language, naive listeners are able to discriminate it from another unknown language after a short excerpt (2 seconds) with 80% accuracy rate [6].

The spoken language represents a complex input carrying several levels of information, both segmental and supra segmental. Perceptual studies are useful to improve our knowledge of the phonological processing of the languages and to provide additional features that may be relevant in automatic language identification. In fact, present systems are mainly based on phonotactic modelling, and, since performances not significantly increase for the last three years, it seems that improvements will require to consider new features (both segmental and supra segmental).

Different approaches in language recognition by human subjects have focused on several parameters as linguistic material, task or subject linguistic background. Experiments focused on the linguistic material treated by the listeners during the task [4], on the evaluation of language proximity [7] or on the function of the previous exposure [8] and/or the level of fluency in foreign languages of the subjects [2].

This paper focuses on the way the mother tongue of the subjects influences the perceptual features they consider in discrimination of five Romance languages (French, Italian, Portuguese, Romanian and Spanish). Four sets of listeners have been selected for the test: French, Romanian, Japanese and American. The next section is an introduction to the Romance language family. Salient features of the five languages studied in this experiment are also provided. The experimental process, and especially the audio material and the sets of subjects are described in Section 3. The results are detailed and interpreted in Section 4.

2. THE ROMANCE LANGUAGE FAMILY
The Romance languages represent a significant part of the today’s most employed languages. They are the exclusive or at least one of the official languages in several regions in the world (Europe, South America, Eastern Canada, Central America etc.). A phonological description of Romance languages points out common genetic features, but also several particularities at segmental and supra segmental level. The structure of the vocalic systems divides the Romance family in two groups, namely languages with two vocalic oppositions (front/back), Italian and Spanish, and languages with three vocalic oppositions (front/back/central or front-rounded), Romanian, French and Portuguese. Romance family admits two of the only four European languages possessing phonological mid and low nasal vowels (French and Portuguese, the other languages being Polish and some Breton dialects) [5]. The consonantal systems are more homogenous in terms of common features, nevertheless specific segments exist (for example, Romanian glottal fricative /h/, Spanish dental fricative /θ/ etc.). The supra segmental level divides languages in four groups according to their syllabic structure: syllable-timed (Italian and Romanian), stress-timed (Portuguese), trailer-timed (Spanish) and fixed-stress or “boundary language” (French) [1].

The present study aims at making arise specific features relevant in Romance language identification. The perceptual experiments have been designed to investigate the accuracy of the discrimination made by human subjects and to evaluate the respective part of the segmental and supra segmental cues in the identification process.
3. EXPERIMENTS

3.1. Audio material

Data of the five languages have been collected in laboratory conditions (22 kHz, quiet environment) and an intensity normalization have been performed. For each language, two speakers (one male and one female) were involved in the learning phase and two other speakers (one male and one female) were considered for the test. Speech utterances consist either in read speech or in story-telling speech.

3.2. Subjects

Four sets of subjects were considered. Two sets correspond with native speakers of a Romance language (French and Romanian), while the subjects of the third and fourth set are Japanese and American English speakers. Each set consists in 20 listeners (male and female subjects), aged from eighteen to sixty, with at least bachelor level. The previous exposure of the subjects to Romance languages is quite homogeneous for each set:

- **French** subjects have studied Spanish at school. None of them were fluent in Spanish or other Romance languages. Additionally, France borders on Spain and Italy.
- **Romanian** subjects have studied French at school. Again, none of them were fluent in French or other Romance languages. Geographically speaking, Romania borders on no other Romance language country. However, fictions produced in South America are commonly broadcast on the Romanian television with the original sound track.
- **Japanese** subjects have not study any Romance language at school, and none of the subjects were previously exposed to Romance languages.
- **American** subjects have studied Spanish and, some of them, French at High School, but none of the subjects were bilingual.

3.3. Experimental conditions

The experimentation was divided in two phases:

- A training phase during which subjects are familiarized with speech samples in each of the five Romance Languages;
- A test phase during which subjects have to decide if stimuli of two excerpts correspond with the same language or not.

After the test completion, subjects are asked for detailing the features they find relevant to discriminate among the languages. During the training phase, two excerpts of 10-second duration are presented to the subjects for each of the five Romance languages. Each stimulus is pronounced by a different speaker (one male speaker and one female speaker per language), and the language sequence is at random. During the test phase, 50 stimuli on an AB model were presented to all groups of subjects: Each stimulus consists in two 6-second utterances separated by a short bell sound. For each stimulus, the subject benefits from a 2-second delay to answer whether A and B are excerpts of the same language or not. No utterance is repeated in the test, and each Li -Lj combination , with \(|i,j| \in [1,…,5]\) is presented twice. Consequently, each different language pair is proposed four times, twice in each order, and each same language pair is presented twice too.

4. RESULTS

4.1. Preliminary significance analysis

The percentages of correct answers given by each set of subjects have been analyzed to evaluate the relevance of the subsequent multi dimensional analysis. Figure 1 displays for each set an histogram of the correct answers according to the language pair involved in the stimuli. Univariate t test have been performed and the figure shows for each language pair if the results are significant (p < 0.001) or not. In most of the cases, the answers provided by the Romanian and the French subjects are significant (with the exception of Portuguese/Portuguese stimuli for both sets and Romanian/Romanian and Romanian/Portuguese for the French subjects). On the contrary, answers from the Japanese subjects are not significant with almost all the language pair: Japanese answer at random. Therefore, a multi dimensional analysis of the Japanese answers is not relevant since no significant trends would be obtained from it. American subjects provide significant results for all items except Portuguese/Portuguese, Portuguese/Romanian, Italian/Italian and Romanian/Portuguese.

4.2. Multidimensional analysis

The Vista software [9] performs a multi dimensional analysis for the French, the Romanian and the American experimental group of subjects. Both same language and different language stimuli are considered in this study.

- **French subjects**

A 2-dimension display of the answers of the French subjects is necessary to take the inter-language differences into consideration. Figure 2 displays the projection of the results into the plane (D1 /D2 ), precisely the first and second axes of the multidimensional analysis. In this plane, languages cluster in three groups (mother tongue, familiar languages and unknown languages). The first dimension splits mother tongue (French) from foreign languages, while the second dimension distinguishes between familiar languages (Italian and Spanish) and unknown languages (Portuguese and Romanian). Furthermore, the third dimension shows a clear separation between the two most familiar foreign languages (Italian and Spanish) while the less familiar languages are still unresolved (Portuguese and Romanian). French subjects seem to be unable to differentiate two unknown Romance languages with only a few seconds of training period. Phonological considerations may also be considered since languages cluster along D2 according to the vowel systems: a three contrast opposition with nasal and central vowels respectively is present in Portuguese and Romanian and not in Spanish and Italian.
Romanian subjects
The analysis of the answers of the Romanian subjects receives also a 2-dimensional solution (Figure 2). It leads us to interpret the first dimension in the same way as for French subjects, since it isolates Romanian (mother tongue) from the rest of Romance languages. The second dimension is interpreted as the complexity of the vocalic system, since languages with front/back oppositions in their vocalic systems (Italian and Spanish) are isolated from languages with more than two oppositions and nasal vowels (French and Portuguese). In addition, a non-represented D1/D3 plane highlights a potential geographical distribution, as it separates the Iberian languages (Spanish and Portuguese) from the rest of Romance languages. Another explanation may be that subjects benefit from the frequent exposure to South American Spanish and Portuguese movies through the Romanian television. That way, D3 splits Romance languages in a familiar language cluster (Spanish and Portuguese) and a less familiar one (French and especially Italian that is not learned at school).

Japanese subjects
Since Japanese subjects have probably answered at random, no multidimensional analysis is performed. However, this result confirms that the previous exposure to the language family is essential in a language discrimination task. An extended learning period is necessary for utterly naive subjects.

American subjects
The same 2-dimensional display was used to explain American subjects’ results. The 1st dimension is associated with acoustical cues and 2 clusters are obtained: Romanian, Portuguese, French vs Spanish, Italian. We interpret this dimension accordingly to the complexity in vocalic systems structure. Languages with two oppositions (Italian, Spanish) are opposed to languages possessing three contrast oppositions in their vocalic systems, the 3rd one being central (Romanian), nasal (French, Portuguese) or front rounded (French). The 2nd dimension distinguishes between familiar languages (French, Spanish) and non familiar languages (Italian, Romanian, Portuguese).

4. DISCUSSION
In this study we focused on the features revealed by the human perception and potentially relevant in automatic identification of five Romance languages. Our findings suggest that naive listeners perceive distinctiveness among unfamiliar languages by employing different strategies in language differentiation. These strategies are mother tongue dependent and are based on several types of information, linguistic and/or extra linguistic. The main linguistic information concerns the complexity of the vocal systems. The sociolinguistic level concerns Romanian subjects who benefit from a special knowledge of Iberian languages via their national television. Furthermore, a previous exposure to the languages is fundamental in the language discrimination task. Truly naive subjects (Japanese listeners) are unable to extract salient patterns after a brief training period.

Further studies should try to perform a more discriminating task by targeting on the vocalic segments of both languages with simple and complex systems. The robustness of linguistic cues determined by experimental methods should be validated by developing a recognition model based on segmental and supra segmental information.

5. REFERENCES
Figure 1: Percentage of correct answers for each set of subjects (French, Romanians, Japanese, Americans). Abscise indicates the language pair (AB in the 1st histogram and AA in the 2nd histogram).

Figure 2: Projection of the multi dimensional analysis of the answers of the French (D1/D2 and D1/D3 display), Romanians (D1/D2 display) and Americans (D1/D2 display).